A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP5878.

Into the Black Box: What Can Machine Learning Offer Environmental Health Research?

Charles W. Schmidt

https://doi.org/10.1289/EHP5878

It was during a 1956 conference that scientists from the RAND Corporation in Santa Monica, California, unveiled what has come to be known as the first artificial intelligence (AI) program. Called the Logic Theory Machine,[1] it could prove complex math theorems by mimicking the problem-solving skills of a human being. This "thinking machine," as its creators described it, reportedly met with an indifferent response, but that's hardly the case with AI now: AI software is embedded in many of the digital devices we live with and use every day, and global spending on the technology neared $36 billion in 2019—an increase of 44% over the previous year.[2]

What is AI? It is not easy to say; the term lacks a consensus definition. "We have trouble analyzing and measuring human intelligence because it is something we experience in our heads," says Sam Adams, a senior AI researcher at RTI International in Research Triangle Park, North Carolina. "So how do we know when we have an artificial version of it?"

Jason Moore, a professor of informatics at the University of Pennsylvania's Perelman School of Medicine in Philadelphia, describes AI as the science of building software and computers that solve problems and reason like humans do. As an example, he cites self-driving cars, which have to discern who else is on the road, read street signs, and make instantaneous decisions on how to maneuver without crashing. AI technology can also enhance human intelligence, as it does when it enables scientists to identify important connections in vast data sets that they cannot detect on their own. Moore says scientists have proposed a newer term, "augmented intelligence," to describe that capability.

Now AI is becoming a powerful research tool in environmental health.[3,4] "I see it as a catalyst for innovation within the environmental health sciences that can help us address many unsolved challenges around how best to utilize large and complex data sets," says Rick Woychik, acting director of the National Institute of Environmental Health Sciences (NIEHS). "Ideally, AI can help us propose new hypotheses or come up with effective solutions for difficult problems."

Environmental health scientists are already using AI to search the literature for useful information, model the effects of pollutants in cells and tissues,[5] and assess air quality on the basis of remote sensing data.[6,7] According to Nicole Kleinstreuer, acting director of the National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods, AI could eventually play a critical role in transcriptomic studies of cells' protein-making machinery and assessments of the "exposome," or totality of an individual's chemical exposures over a lifetime.

Still, experts point out that AI can also generate misleading results when used incorrectly. AI algorithms can be difficult to train, and many of them are "black boxes"—meaning their internal calculations are either proprietary information or too complex for people to understand.[8] Scientists may justifiably wonder if a black box will behave as expected when it processes real-world data or if it will pick up on confounding signals that compromise its predictions.

Ivan Rusyn, a professor of toxicology at Texas A&M University, cautions that some scientists may oversell the technology and mislead the general public by suggesting that AI-enabled solutions to difficult problems in medicine and environmental health are "just around the corner and within reach."

Moore agrees, adding that when it comes to AI in environmental health, slow and steady is the way to go. "We want to be enthusiastic while tempering expectations as scientists identify the right approach to each problem," he says.

A Machine-Learning Tutorial

The driving force behind AI is machine learning, which refers to how computer algorithms improve at performing assigned tasks with increasing experience.[9] One way they do that is by learning to recognize patterns in data. Training in pattern recognition can be either supervised (coached by humans) or unsupervised, meaning the algorithms are turned loose on data to identify patterns on their own.

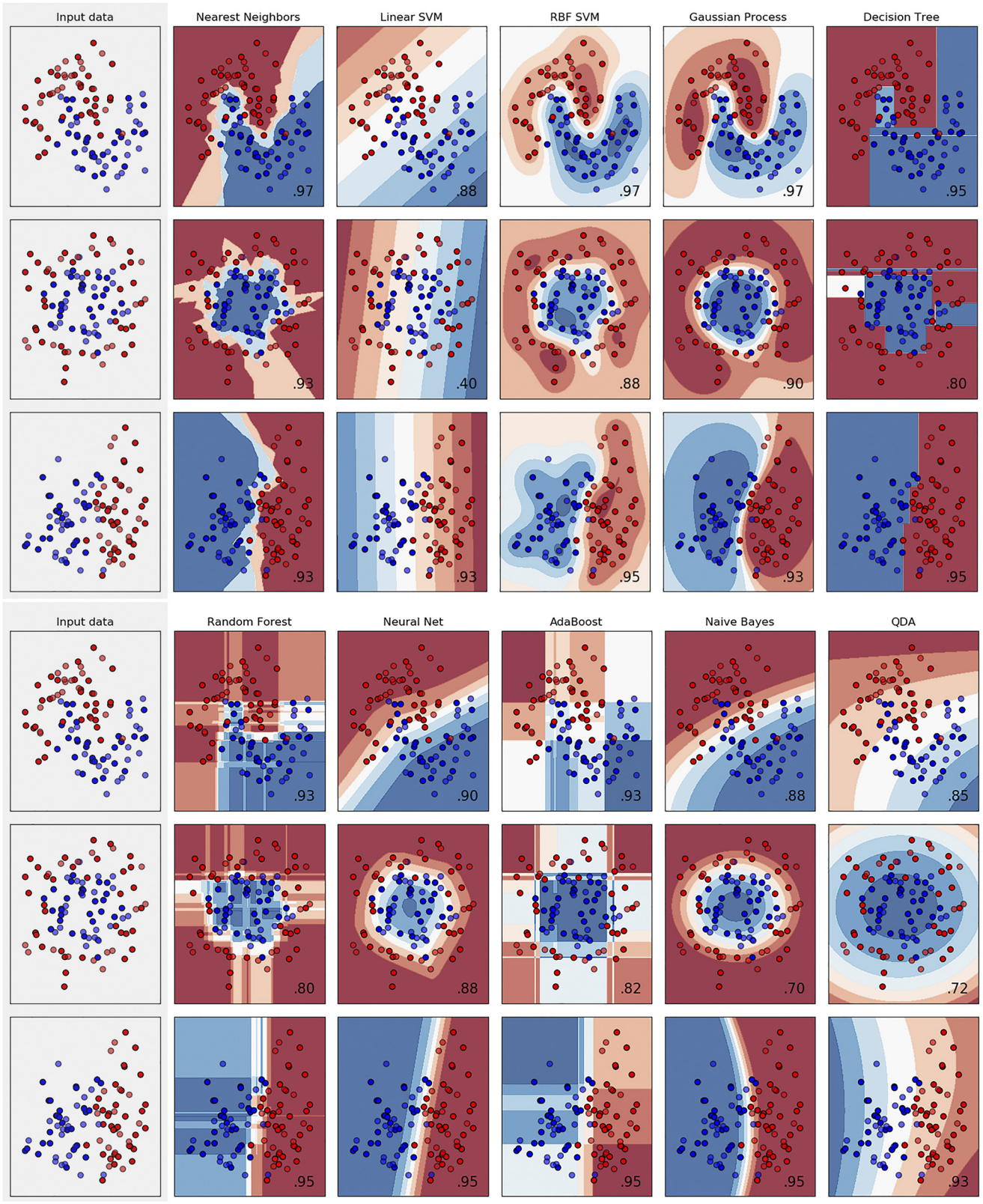
Supervised algorithms must first be trained with labeled data sets that show them how to recognize, for instance, a cat in a digital photo, a gene in a DNA sequence, or the likely price of a home in a given neighborhood. Depending on the underlying nature of the data, the algorithms' predictions arrive in one of two categories: either a discrete classification (such as "cat" or "gene") or a regression category in which the prediction describes a measured value within a continuous variable (such as "price").

Unsupervised algorithms self-organize data without any such guidance. With a common technique called cluster analysis, for instance, these algorithms automatically sort data with similar features into groups. Because scientists might not know to look for those data groupings in advance, cluster analysis can lead to new and unanticipated discoveries.

An even more powerful subclass of machine learning, called deep learning, relies on layers of algorithms that are arranged to mimic the architecture of the human brain.[10] Convolutional neural networks (CNNs), for instance, are deep-learning models inspired by the arrangement and functioning of the human visual system. CNNs are at the core of most computer vision applications today—such as Facebook's automated photo-tagging system or the interpretation of remote sensing data.

But there are many other types of deep-learning models. A recurrent neural network is a model that's particularly good at finding patterns in time-series data, meaning data sets that change over time (think stock market prices or fluctuations in ozone concentrations during the day). Yet another kind of deep-learning model, called an autoencoder, is used for unsupervised machine learning and can be applied to reconstruct complete digital images and other data representations from minimal sets of key information. In some cases, they are used to filter out extraneous "noise," which is useful for sharpening digital images.

Selecting the right model for the job is crucially important, though it is not always obvious which one to pick. "One of the most common questions I get is, 'What sort of model should I use with my data?' " says Marianthi-Anna Kioumourtzoglou, an assistant professor at Columbia University's Mailman School of Public Health who uses AI in health studies of chemical mixtures. The answer, she says, is that researchers should start by clearly framing the question they want to answer.

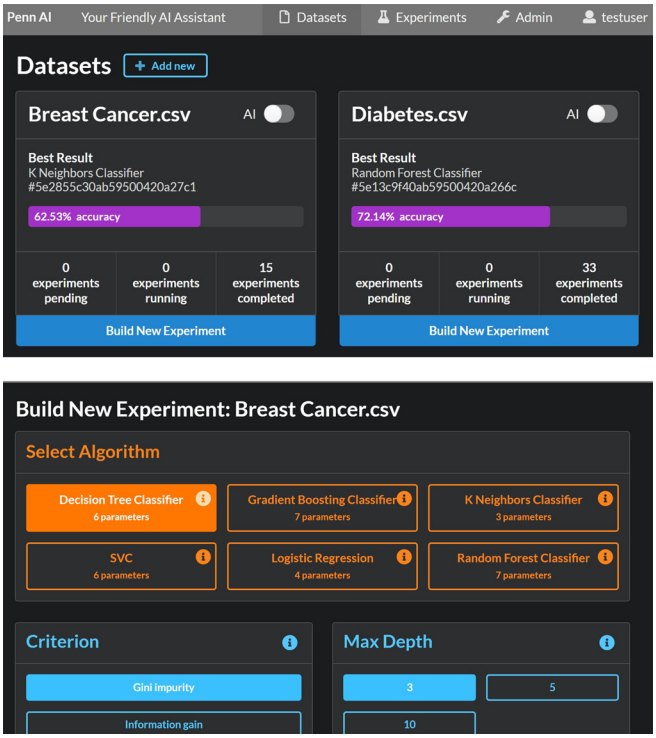


This figure shows three data sets (rows) with two variables (x- and y-axes) and two outcomes (blue or red). Each of the 10 machine-learning methods (columns) attempts to classify dots as blue or red by building mathematical functions of the x and y variables. The shades of color reflect the confidence of the model in classifying each dot as blue or red. The numbers represent the classification accuracy or the proportion of dots correctly assigned a blue or red outcome by the model. Each method detects different patterns and performs differently on each data set. One of the challenges of machine learning is knowing which method is the best choice for a specific data set. Image: 3-Clause BSD License. Figure created using the Scikit-learn library.[18]

David Carlson, an assistant professor of civil and environmental engineering at Duke University, says the risk to avoid is "overfitting," or the tendency of a poorly chosen model to capture noise in the data instead of real information. In these cases, the model will generate unreliable predictions, whereas well-chosen models, he explains, will be generalizable. In other words, a well-chosen model will be capable of adapting properly to new data that it has never seen before. Scientists can apply several statistical tests to validate their models so they can be more confident in the models' generalizability.

Although model selection involves specialized skills in statistics and computer science, some researchers are also turning to a growing number of open-source software packages that will fit models to their data automatically. At a 2019 conference devoted to AI for environmental health,[11] UPenn's Moore described one such package called PennAI, which was developed by his own research team. "You just load your data set, push a button, and the AI takes over and launches what it thinks is the best model to run," he said. According to Moore, PennAI is able to do this because it creates a knowledge base of which models tend to work with different kinds of data, similar to the systems that commercial entities such as Amazon use to suggest items you might want to buy based on your shopping history.

In other words, Carlson explains, PennAI and many other packages are intended to make AI tools more accessible to a broad audience. But, he adds, while such tools *are* more accessible than ever, "I personally do not think that such systems are there yet, and you need a lot of expertise and understanding to use and correctly interpret the output of such a system."



The open-source PennAI software aims to simplify machine learning for the user. On the launch screen (top), users choose from available data sets to perform analyses. The "Best Results" box shows which algorithm has performed most accurately on each data set. The user can also browse all the results for each data set by clicking on the "experiments completed" box. From here, the user has the option of toggling to the "AI" option to let the software automatically choose an appropriate machine-learning algorithm and parameters. Or, on the "Build New Experiment" screen (bottom), the user can manually select algorithms and parameter settings. Image: Courtesy Jason Moore.

The Artificial Intelligence Landscape in Environmental Health Today

As AI moves into environmental health research, near-term opportunities for the technology are arising on several fronts. Text analytics (also known as text mining) uses machine-learning algorithms to extract useful information from papers and reports. This "is a big area of interest for us," says Jerry Blancato, director of the Office of Science and Information Management at the U.S. Environmental Protection Agency (EPA). Blancato says text analytics will ideally allow for better ways to manage, query, and categorize data from different sources.

According to Paul Whaley, a research fellow at Lancaster University in the United Kingdom and the Evidence-Based Toxicology Collaboration in the United States, advances in text analytics will go a long way toward boosting the efficiency of systematic review, a highly methodical process by which scientists collect evidence from multiple sources that can help answer certain questions. As it stands now, systematic reviews rely heavily on research associates who have to read through hundreds or even thousands of documents. Whaley says the EPA and the NIEHS have both begun to automate these initial screenings with machine-learning algorithms that classify the documents according to keywords in titles or abstracts.

More complex text analytics may eventually allow algorithms to read and comprehend entire sentences, although these programs do not yet have the necessary rich, granular understanding of language. "That's the sort of capability we're really looking for," Whaley says. "Classifications are helpful, but more than that, we need machine-learning systems that can read through the reports and extract relevant information for us. That way, instead of manually extracting data from, say, twenty-five reports, you could automatically pull it from thousands of potentially useful documents and wind up with much larger, richer data sets than can be assembled manually."

Whaley adds that an important step in that direction would be to assemble a "full-text corpus" of annotated studies that could be used to train algorithms to read technical language more effectively. A full-text corpus is a set of documents within which the important information has been highlighted or tagged by hand. According to Whaley, algorithms trained on such a knowledge base will learn to identify and extract similar information when they are exposed to it to in other documents later.

At the NTP, researchers are using analogous methods with an eye toward developing computerized systems for predicting chemical toxicity. Toward that end, Kleinstreuer's group and researchers at Oak Ridge National Laboratory are jointly developing algorithms that, as a first step, will identify high-quality papers in the toxicology literature. During this initial process, reviewers have to read the studies and then extract information on, for instance, the protocols, types of chemicals tested, and observed effects. The aim is to use the information in those papers as source material for databases that relate chemical structures to toxic end points such as mortality, endocrine disruption, and protein reactivity, among others. In turn, these databases can be used to train machine-learning models used by other teams investigating chemical safety.

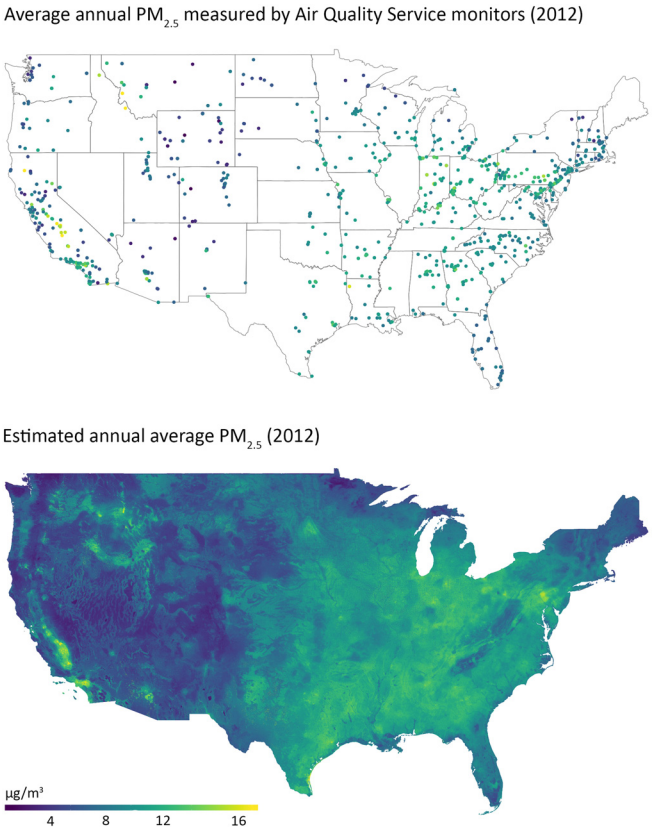
Assembling the databases requires that NTP researchers put the published information into machine-readable formats that computer algorithms can work with. "A lot of what we're doing now is brute force curation to digitize studies that are not computationally accessible," Kleinstreuer says. She adds that NTP researchers recently curated a database of rodent LD_{50} values (which describe the dose that kills 50% of a group of exposed animals) associated with approximately 15,000 chemical structures. Kleinstreuer says that as model development evolves, the entire process—from selecting papers, to curating databases, to finally developing algorithms that

predict toxic effects from exposure to untested chemicals—could in time be accomplished with AI.

Applying machine learning to field- and satellite-based remote sensing data is yet another emerging development. At the EPA, scientists are using the technology to map floodplains and mosquito habitats and to develop predictive models that warn of toxic algal blooms. Elsewhere, other researchers are using it to estimate air pollution levels. One of these scientists is Scott Weichenthal, an epidemiologist at McGill University. During a recent project, Weichenthal's team found that when applied to satellite imagery, CNNs predicted concentrations of fine particulate matter ($PM_{2.5}$) with nearly the same accuracy as a model used by the World Health Organization (WHO) to assess air quality for its Global Burden of Disease study.[12]

The WHO's model, which is called the Data Integration Model for Air Quality, relies on many different inputs, such as chemical transport features and pollution measurements gathered from sensors on the ground. Weichenthal and his colleagues trained their model by pairing ground-level sensor data from approximately 6,000 sites in 98 countries with corresponding satellite data for each sensor location. Once trained, the model could predict variation in $PM_{2.5}$ levels solely based on land-based features, "and all you need to run it is the satellite picture," Weichenthal says.

Building on this approach, Francesca Dominici, a biostatistician and codirector of the Data Science Initiative at Harvard University, has related machine learning–derived estimates of airborne $PM_{2.5}$ concentrations to changes in mortality among older Americans.[13] For that effort, she and her colleagues relied on a model[14] that combined ground- and satellite-based measures and applied machine-learning algorithms to the data to estimate pollution levels at the square-kilometer level throughout the United States. They paired the predicted values with data from millions of Medicare claims, gathered from each zone between 2000 and 2012. Their analysis indicated that increases of 10 $\mu g/m^3$ in $PM_{2.5}$ and 10 ppb in ozone were associated with increases in all-cause mortality of 7.3% and 1.1%, respectively.[13]



Investigators including Frederica Dominici developed a machine-learning model to predict $PM_{2.5}$ concentrations across the United States. The model incorporates remotely sensed data, estimates of ground-level $PM_{2.5}$ and total atmospheric aerosols, meteorological data, land use data, and more. The training set (top) was based on monitoring data from the U.S. Environmental Protection Agency's Air Quality System. The model produced an image (bottom) that closely mirrors the ground-truthed data but offers a finer spatial scale. Image: Courtesy Benjamin M. Sabath.

Issues of Trust

Still, Dominici describes the modeled $PM_{2.5}$ predictions as guesses, adding, "We're not there yet in terms of quantifying how good the guesses from machine learning are." That is especially true when the predictions come from black boxes that, as she says, breed uncertainties "that we cannot afford to ignore when we're estimating health effects."

Weichenthal agrees that the technology isn't without its shortcomings. He acknowledges that the estimates in his work became increasingly unreliable outside the areas where the model was initially trained. Moreover, given that the model's internal calculations are somewhat opaque, the specific features of the built environment that drive its predictions are not known.

In an especially egregious circumstance of poor guesswork that came to light during the 2018 California wildfires, Google used a proprietary black-box machine-learning algorithm from another company to power its search page weather widget. The widget claimed air pollution levels were safe,[15] even as people in the area were watching the ash build up on their cars.[8] According to Carlson, computer scientists are currently experimenting with ways to open up deep neural networks and other black boxes to expose their internal calculations or to produce interpretable models with comparable accuracy.

Meanwhile, any model's accuracy depends, in large part, on the quantity and quality of the data to which it is exposed and differences between training data and real-world data. Carlson wrote in 2019 that these accuracy-altering differences "can cause significant problems for machine-learning methods."[8] Carlson claimed that "modifying a single pixel can completely alter an algorithm's understanding of an image" and a small decal stuck to a stop sign "can fool even a modern industrial computer vision system for self-driving vehicles."[8]

Making sure that machine-learning algorithms used in environmental health have sufficient access to high-quality data is now a priority for the field. "Nothing in AI is going to work if you do not pay attention to data quality," says Woychik, adding that the NIEHS is highly focused on developing sustainable systems for generating data that can be easily shared with researchers around the world. Fundamental to that goal, he says, is that data production abides by the FAIR Guiding Principles, which were first published in 2016.[16] Those principles state that data, and associated data objects such as code, should be findable, accessible, interoperable, and reusable by humans and machines alike.

Toward that end, the NIEHS is currently overhauling its cyberinfrastructure to better prepare for AI uses. The institute has hired new staff tasked with assembling a plan for cyberinfrastructure management, including better ways to collect, annotate, and archive data for ongoing and future use.[17] Once those systems are in place, "we can think about ways to do more complex experiments with AI," Woychik says, "but without overpromising on the potential when so much is still speculation."

Similarly, officials at the EPA recently established a formal steering committee that's become a gathering point for people interested in AI who want to provide training, advice, or consultation. "We have many people with deep expertise, and we're

looking to share the wealth and build up collaboratives," says the EPA's Blancato.

Adams at RTI agrees that most of the current environmental health focus is still on preparing data for use by machine-learning algorithms. "Facebook and other companies are successful [in] doing this because they are working with terabytes of data," he says. "The rest of us doing science are still investing resources to label data and make it available for people to use. And what we can do with the technology [depends on] how well we integrate the data we collect."

Charles W. Schmidt, MS, is an award-winning journalist in Portland, Maine, whose work has appeared in *Scientific American*, *Nature*, *Science*, *Discover Magazine*, *Undark*, the *Washington Post*, and many other publications.

References

1. Newell A, Simon HA. 1956. *The Logic Theory Machine: A Complex Information Processing System. P-868*. Santa Monica, CA: The RAND Corporation.
2. Shirer M, D'Aquila M. 2019. Worldwide spending on artificial intelligence systems will grow to nearly $35.8 billion in 2019, according to new IDC spending guide. [Press release.] International Data Corporation, 11 March 2019. https://www.idc.com/getdoc.jsp?containerId=prUS44911419 [accessed 3 February 2020].
3. Research Triangle Environmental Health Collaborative. 2019. *11th Environmental Health Summit, Artificial Intelligence in Environmental Health Science and Decision-Making*. 18–19 October 2018. Research Triangle Park, NC: North Carolina Biotechnology Center.
4. Miller TH, Gallidabino MD, MacRae JI, Hogstrand C, Bury NR, Barron LP, et al. 2018. Machine learning for environmental toxicology: a call for integration and innovation. Environ Sci Technol 52(22):12953–12955, PMID: 30338686, https://doi.org/10.1021/acs.est.8b05382.
5. Luechtefeld T, Marsh D, Rowlands C, Hartung T. 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicol Sci 165(1):198–212, PMID: 30007363, https://doi.org/10.1093/toxsci/kfy152.
6. Hong KY, Pinheiro PO, Minet L, Hatzopoulou M, Weichenthal S. 2019. Extending the spatial scale of land use regression models for ambient ultrafine particles using satellite images and deep convolutional neural networks. Environ Res 176:108513, PMID: 31185385, https://doi.org/10.1016/j.envres.2019.05.044.
7. Weichenthal S, Hatzopoulou M, Brauer M. 2019. A picture tells a thousand...exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. Environ Int 122:3–10, PMID: 30473381, https://doi.org/10.1016/j.envint.2018.11.042.
8. Rudin C, Carlson D. 2019. The secrets of machine learning: ten things you wish you had known earlier to be more effective at data analysis. In: *Operations Research & Management Science in the Age of Analytics*. Netessine S, ed. Catonsville, MD: The Institute for Operations Research and the Management Sciences, 44–72, https://doi.org/10.1287/educ.2019.0200.
9. Meserole C. 2014. What Is Machine Learning? Brookings Institute. https://www.brookings.edu/research/what-is-machine-learning/ [accessed 3 February 2020].
10. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature 521(7553):436–444, PMID: 26017442, https://doi.org/10.1038/nature14539.
11. National Academies of Sciences, Engineering, and Medicine. 2019. *Leveraging Artificial Intelligence and Machine Learning to Advance Environmental Health Research and Decisions: Proceedings of a Workshop—in Brief*. Washington, DC: National Academies Press. https://www.nap.edu/catalog/25520/leveraging-artificial-intelligence-and-machine-learning-to-advance-environmental-health-research-and-decisions [accessed 3 February 2020].
12. Hong KY, Pinheiro PO, Weichenthal S. 2019. Predicting Global Variations in Outdoor PM2.5 Concentrations Using Satellite Images and Deep Convolutional Neural Networks. https://arXiv.org/abs/1906.03975 [accessed 3 February 2020].
13. Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Choirat C, et al. 2017. Air pollution and mortality in the Medicare population. N Engl J Med 376(26):2513–2522, PMID: 28657878, https://doi.org/10.1056/NEJMoa1702747.
14. Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang Y, Schwartz J. 2016. Assessing $PM_{2.5}$ exposures with high spatiotemporal resolution across the continental United States. Environ Sci Technol 50(9):4712–4721, PMID: 27023334, https://doi.org/10.1021/acs.est.5b06121.
15. McGough M. 2018. How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say. *Sacramento Bee*, California section, online edition. 7 August 2018. https://www.sacbee.com/news/california/fires/article216227775.html [accessed 10 January 2020].
16. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A. 2019. Addendum: the FAIR Guiding Principles for scientific data management and stewardship. Sci Data 6(1):6, PMID: 30890711, https://doi.org/10.1038/s41597-019-0009-6.
17. National Institute of Environmental Health Sciences. 2019. Informatics and Information Technology Strategic Roadmap, Fiscal Years 2019–2021. https://www.niehs.nih.gov/about/informatics-it/index.cfm [accessed 18 November 2019].
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. 2011. Scikit-learn: machine learning in Python. J Machine Learning Res 12:2825–2830.